CICE Hiroshima University, *Journal of International Cooperation in Education*, Vol.9, No.1, (2006) pp.45 ～ 57

Science-based Self Evaluation of Learning at the World Bank[1]

Marlaine E. Lockheed
Formerly World Bank

Abstract

Scientific rigor is being re-introduced into evaluation research after some period of absence, with strong proponents of science in evaluation found among economists. This paper describes efforts to introduce greater scientific rigor into self-evaluation of learning programs of the World Bank. These programs are designed for two sets of learners: (a) World Bank staff, and (b) government officials and policy makers, technical experts, business and community leaders and civil society stakeholders from developing countries. Efforts to introduce science-based evaluation methods discussed in this paper include use of: mixed methods, valid and reliable measures of learning, counterfactuals, scientific sampling, and multivariate statistical analysis. The paper describes these efforts and concludes by reviewing the evidence from these evaluation about the features of effective learning programs for adult learners

Introduction

Scientific rigor is being re-introduced into evaluation research after some period of absence, with strong proponents of science in evaluation found among economists (Kremer et al 2002). Most science-based evaluations are carried out by independent evaluators, defined by OECD as entities and persons free of the control of those responsible for the design and implementation of the development intervention" (OECD 2002). By comparison, self-evaluation is defined by the OECD as "an evaluation by those who are entrusted with the design and delivery of a development intervention." Institutional research and evaluation, therefore, falls into the category of self-evaluations, since—at some level—institutional evaluation offices of necessity report to those who are also responsible for the implementation of the institution's programs. However the introduction of science-based methods for self-evaluation improves its validity, transparency and ultimately independence.

This paper describes efforts to introduce greater scientific rigor into self evaluation of learning, as it has been carried out within the World Bank for programs directed at both World Bank staff and the clients of the World Bank Institute (WBI). WBI supports the delivery of about 800 learning and capacity enhancement activities to approximately 56,000 people in 100 countries annually. These programs cover 16 different sectoral and thematic areas

[1] Until her retirement in 2004, Marlaine E. Lockheed was the head of the World Bank Institute Evaluation Group. Helpful comments on an earlier draft were received from Henry Braun (Educational Testing Service) and Patrick Grasso (OED). The findings, interpretations, and conclusions expressed in this paper are entirely those of the author and do not necessarily represent the view of the World Bank Group.

plus global learning related to Knowledge for Development, Governance and Monitoring and Evaluation. In addition, the Learning Board of the World Bank is responsible for internal staff learning through over 2000 learning activities annually delivered to approximately 10,000 Bank staff in Washington and country offices. These programs cover professional and technical training in the sectors, behavioral and managerial learning programs, and various program offered through regional offices.

From FY02 through FY04, an intensive effort was made to introduce science-based methods into the evaluation of these programs, in five areas: (a) using mixed methods, (b) measuring learning, (c) establishing counterfactuals, (d) sampling, and (e) analysis. The paper describes these efforts and concludes by reviewing the evidence from these evaluations about the features of effective learning programs for adult learners.

Mixed Methods

For many years, evaluation experts have recommended the use of mixed methods of evaluation to "triangulate" findings as well as to expand the types of evaluation questions that are addressed (Campbell & Stanley 1963; Denzin & Lincoln 1994; Scriven 1991; Rossi & Freeman 1993). While previous evaluation work on learning at the World Bank had used a variety of methods, each evaluation typically used only a single approach: surveys, interviews or focus groups, although there were exceptions (Bussman, West Meier & Hadorn 2001; Universalia 2001). Beginning in 2001, however, all major evaluations began using mixed methods, including: surveys, interviews, focus groups, product assessments, retrieval of archival data, document review, observations of training, and tests. As a consequence, the share of evaluations using mixed methods increased sharply. Fifty-three percent of evaluations initiated in 2001 or later and published in 2003 and 2004[2] used three or more methods, compared with 27 percent initiated earlier and published in 2001 and 2002. Fifty-four percent of earlier evaluations used one method only, compared with only six percent of later evaluations (Table 1). In addition to mixed methods, the new evaluations began collecting information from a variety of informants, including: participants, colleagues, subordinates, supervisors, learning providers, client counterparts, independent reviewers. This enabled the evaluators to analyze and interpret the same phenomena from several points of view.

Measurement of Learning

From FY97 through FY01, the learning outcomes of WBI programs were measured in two ways: self reports of perceived learning gains and short multiple-choice tests of learning, typically administered both before and after the learning event, so as to measure gains. No evidence was available regarding either the degree to which the pre- and post-tests could be considered parallel forms of the same test or the extent to which the tests were internally consistent.

[2] This review covers evaluations published or available in final draft as of May 2004.

Table 1. Percent of Evaluations Using Multiple Methods by Publication Year

	Number of Evaluation Methods Used		
Date of publication	One	Two	Three or more
2001 and 2002 (N = 11)	54%	18%	27%
2003 and 2004 (N = 17)	6%	41%	53%

Source: author's analysis

Parallel forms

Prior to FY02, two test forms were prepared from a single item pool from which items were randomly assigned to either the pre- or post-test form (Eckert 2000). A total of 68 such tests were developed from 1999 to 2001 (Ouchi, Shi & Zhou 2001). The random assignment of items to tests did not ensure that each test included items that addressed the various topics of the learning event, or that item difficulties were similar for the two forms. To improve the comparability of the pre- and post-tests, in FY02 the evaluators introduced the concept of the test specification matrix as a blueprint for developing test questions. This was a two-dimensional matrix, covering only content topics and estimated level of difficulty (Le Rouzic & Shi 2003). While parallelism improved as both pre- and post-tests included items from the cells of the matrix, an item-by-item comparison of difficulty of pre-tests and post-tests in 2002 and 2003 suggested that the tests were still not fully parallel forms of the same test, but rather different tests; the median correlation coefficient for item difficulties for "matched" items on 17 tests was only .02 (Le Rouzic& Shi 2003). However, correlations between pre- and post-forms were high, and the internal consistency of the tests had improved, enabling the use of the pretest to control for participant intake knowledge, in multivariate analyses of the effects of the learning programs. In 2003, the evaluators issued guidelines for pairing items to improve the parallelism of pre- and posttests at the item level.

At about the same time, the evaluators began to develop a certification test for use with World Bank staff. Again, multiple forms were developed, but through a procedure that ensured a high degree of parallelism among forms. Each test-taker's form was generated through a stratified random sampling process from a pool of items; the strata were learning-module based and designed to maintain equivalence across all exam forms that were generated through the sampling procedure (Khattri, Shi, Palmisano & Echternacht 2003).

Reliability

Analyses of the internal consistency reliability of tests administered FY99-FY01 showed levels for Cronbach's alpha that were well below professional standards: the average reliability coefficient was about .44 with a range of .10 to .70 (Cronbach, 1951; Ouchi, Shi & Zhou 2001). In addition to introducing the test specification matrix, the evaluators worked with the training providers to improve the quality of test questions, using statistical item analysis

of previous tests to clarify items, increasing the number of items to make the test more representative of the course content, providing test development guidance and feedback to the training providers, and in general helping with basic test development expertise. With this help, the internal consistency reliability improved, to an average of .56 in FY02 and a range from .29 to .75, while tests developed without this assistance retained a low internal consistency reliability of .41 (Le Rouzic & Shi 2003).

On the staff learning side, the internal consistency reliability of the certification test was assessed at .83, meeting professional standards (Khattri, Shi, Palmisano, & Fein 2003). Tests for other staff learning courses were developed using the methods newly established for WBI learning programs, resulting in professional standards for reliability, which averaged .82 with a range of .77 to .88 (Le Rouzic 2003).

Establishing Counterfactuals

WBI and the Learning Board were not accustomed to evaluations that could actually assess the effectiveness (or non-effectiveness) of their learning program through the use of counterfactuals or "control" groups. For the new evaluations, two types of counterfactuals were used: (a) post-learning matching of participants with non-participants using either propensity-score matching or other types of trait-by-trait matching techniques, and (b) pre-learning matching of participants with similar non-participants scheduled for subsequent participation (Rubin & Thomas 1992; Campbell & Stanley 1963).[3] In no case were randomized trials utilized, and counterfactual evaluation methodology was not applicable in all cases. Three "post-learning" and two "pre-learning" matching examples follow, all from staff learning.

Post-learning Case I: A five-day course for support staff

This evaluation used four different comparisons: (a) participants and non-participants matched by propensity score, (b) participants at time t1 compared with participants at time t2, (c) participants and non-participants matched by pre-training performance evaluations, and (d) instrumental variables estimation. The propensity score matching method used the Bank's internal personnel database to identify two groups of staff: those who had taken the course and comparable staff who had not (Bardini, Gunnarsson & Palmisano 2003).[4]

Post-learning Case II: A professional conference

This evaluation also used the Bank's internal personnel database to establish two groups of staff: those who had participated in a large conference-type event and those who had not. Three propensity score matching procedures were used: (a) nearest neighbor matching with replacement and equal weight of forward and backward matches, (b) nearest neighbor

[3] Propensity score matching estimates the probability of an individual's participation in a learning event; it is possible to do when data exist for a large number of individuals, only some of whom have participated.
[4] The identify of staff in each group was kept confidential.

matching with replacement and random draw of forward and backwards matches, and (b) radius matching. Archival data on staff in each group were analyzed (Eckert, Palmisano, Gunnarsson & McIntosh-Alberts 2004). In addition, a small number of participants and non-participants selected informally were interviewed.

Post-learning Case III: On the job team learning

This evaluation compared the products of teams that had participated in a training program with the products of teams that had not participated in this program. After three years, half of the teams had produced products, which were compared with equivalent products of teams not participating in the program, on ratings provided by blind reviewers (Quizon, Ouchi & Gunnarsson 2004).

Pre-learning Case I: An introductory course on World Bank operations

This evaluation used two types of comparisons: (a) participants compared with themselves at two points in time[5], and (b) participants compared with similar staff who had not participated in the course (Liu 2003; Le Rouzic 2003). This design may be indicated in this fashion[6]:

Group 1 O X O O
Group 2 O

Pre-learning Case II: A program for managers

A three-module course was offered with different cohorts of participants, all of whom were senior Bank staff. A complex evaluation design was used, whereby participants served as their own controls through a pre-test and were also compared with similar participants in a subsequent cohort. Data were collected before and three months after the learning events. As part of the data collection, both participants and their subordinates were surveyed (Zia 2004). This design may be indicated as[7]:

Group 1 O X O
Group 2 O X O

Sampling

Previously, total populations of participants in specific courses were contacted for purposes of evaluation. As the reach of both WBI and staff learning programs grew, from hundreds of courses and participants to thousands and tens of thousands, this evaluation approach became unfeasible. Instead, evaluators turned to sampling; four examples follow.

[5] In the figure, O refers to “observation” or data collection and X refers to the training program. The one-group pretest-posttest design described by Campbell and Stanley (1963)
[6] The static-group comparison design of Campbell and Stanley (1963)
[7] Similar to the recurrent institutional cycle design of Campbell and Stanley (1963)

Sampling Case I: Evaluation of staff learning

In FY03, a random sample of 242 staff learning courses were selected for evaluation, representing about ten percent of all staff learning and nearly 50 percent of unique activities offered. To minimize a potential selection bias, courses were randomly selected from those meeting four criteria: courses appeared in the Bank Learning Catalogue at least two weeks prior to delivery, Bank staff comprised at least half of the attendees, the course was not an e-learning course, and the course had not been previously evaluated in FY03 (Chard & Arango 2003).

Sampling Case II: Evaluation of six sectoral and thematic programs of WBI[8]

A two-stage nested sampling procedure was used. At the first stage, a sample of learning activities were selected from each of the six programs being evaluated, stratified according to mode of delivery (face-to-face or videoconferencing). Next from within each sampled learning activity, about 50 participants were randomly sampled for a total sample of 1225 participants (Khattri et al. 2002).

Sampling Case III: A country-focused evaluation

A major evaluation of the effects of WBI programs in five countries included total populations of participants, where these numbers were smaller than about 200 and randomly sampled participants from larger country participant cohorts, for a total of about 1200 cases. This sampling procedure introduced issues of weighting in subsequent analyses (Bardini, Gunnarsson, Manjieva & Narozhnaya 2003; Eckert, Sousa & Gunnarsson 2004; Khattri, Bachrach & Jiang 2003; Quizon & Chard 2003; Zia, Al-Sayyid, Tawila & Gunnarsson 2003; Quizon, Chard & Lockheed 2004).

Sampling Case IV: A second sectoral and thematic evaluation

A second evaluation of WBI thematic programs adopted a sampling strategy that both provides information at the country level, but also includes sufficient numbers of participants to evaluate the program as a whole. A three-stage sampling strategy was employed: (a) a list of all eligible program participants who had at least one piece of contact information was prepared, to which was added all eligible participants being surveyed for a parallel exercise being carried out in seven countries, (b) these participants were stratified by country, and (c) a 30-50 percent random sample of program participants was drawn from those countries with high proportions of participants in one or more of the four programs, yielding a total of about 800 participants equally distributed across the four programs (Liu et al. 2004).

[8] WBI organizes its learning and capacity enhancement offerings according to “themes” (e.g. Education) which are largely grouped into sectoral (e.g. Human Development) units.

Analysis

Previously, evaluation analyses concentrated on (a) descriptive statistics for single learning events, (b) summaries of informal interviews and focus groups, and (c) assessments of learning (e.g. Prom-Jackson, Cooper, Martin, Kategile, Palmisano & Arango 2002; Prom-Jackson, Cooper, Palmisano, Latib, Rickwood & Arango 2002). With the use of counterfactuals, simple tests of program effectiveness could be utilized. With better measurement instruments and scientific sampling, multivariate analyses testing hypothetical models of the determinants of effectiveness could identify features of more effective activities. These two new analytic approaches to institutional self-evaluation are discussed below.

Tests of program effectiveness

The use of propensity score matching and other matching techniques allowed for simple t-tests of program effects. For the first time, it was possible to assess the value-added (if any) of the course or program with respect to specific outcomes. In addition, it was possible to compare the effectiveness of the course or program as perceived by participants with its effectiveness in comparison with a control group.

The results, summarized in Table 2, are relatively consistent. Staff performance as rated by managers of participants[9] was not significantly different from that of non-participants. However, independently evaluated measures of performance—scores on tests[10] and ratings by blind observers[11]—were more sensitive to changes caused by courses. And in one case, the sustained use of the types of knowledge imparted by the course was greater for participants than for matched non-participants. In general, interviews and focus groups with participant found positive assessment of these courses for relevance, usefulness and application. Interviews of participants and non-participants also revealed differences in their views.

Design effects do not account for these differences. The differences in test scores were not only statistically significant, they were also meaningful, with effect sizes .25 or more, while the lack of statistical significant in managers' ratings was not due to small sample sizes. Details are provided in Annex A.

Models identifying features of effective learning programs

The evaluations moved from simple bivariate descriptions and OLS models that estimated the effects of various activity features, net of participant intake characteristics, to two-stage regression models and multi-level models. The results of these evaluations show remarkable consistency regarding the features of effective learning programs for adults. Table 3 summarizes the findings from five meta-analyses of these studies:

[9] On a five-point scale
[10] In percentages
[11] On a six-point scale

Table 2. Differences in Outcome Measures for Program Participants Compared with Similar Non-participants

Program/ Course	Outcome measure	Score		
		Participant	Non-Participant	Level of Sig.
Pre-Learning Case I	Test of knowledge retained	61.0	46.2	.01
	Use of general knowledge	65.7	57.2	.05
	Use of specific knowledge	63.3	56.7	.10
Post-Learning Case I	Assessment of behavior by manager	3.68	3.70	n.s.
Post Learning Case II	Assessment of performance by manager	3.68	3.67	n.s.
	Assessment of behavior by manager	3.66	3.66	n.s.
Post-Learning Case III	Blind rating of team product	5.75	5.50	.05

Source: Bardini, Gunnarsson & Palmisano (2003); Liu (2003); Quizon, Ouichi & Gunnarssoon (2004); Eckert, Palmisano, Gunnarsson & McIntosh-Alberts (2004).

Table 3. Features of Adult Learning Programs Associated with Higher Quality and Effectiveness, Various Studies

	Outcome Measures		
Learning Activity Feature	Perceived quality	Perceived effectiveness/Usefulness	Learning/ Perceived earning
Activity Quality			
Longer duration	Ouchi and Le Rouzic	Khattri et al.	Le Rouzic and Shi; Chard and Arango
Participatory/Action Planning	Chard and Arango	Chard and Arango; Khattri et al.; Quizon, Chard and Lockheed	-----
Professionally designed (w/partner)	Ouchi and Le Rouzic	Quizon, Chard and Lockheed	-----
Program follow-up/Part of series	-----	Khattri et al.; Quizon, Chard and Lockheed	-----
Activity Tailoring			
Participant homogeneity	Ouchi and Le Rouzic	Quizon, Chard and Lockheed	Le Rouzic and Shi
Aligned with participants' job	Ouchi and Le Rouzic	-----	Chard and Arango
Country focus	-----	Quizon, Chard and Lockheed	-----

Source: Chard & Arango (2003); Khattri et al. (2002); Le Rouzic & Shi (2003); Ouchi & Le Rouzic (2002); Quizon, Chard & Lockheed (2004).

(1) Chard and Arango, 2003, covering 242 World Bank staff learning activities assessed by 3,326 participants;
(2) Khattri et al. 2002, covering 25 WBI client learning activities assessed by 1,225 participants;
(3) Le Rouizic and Shi, 2003, covering 28 WBI client learning activities assessed by 1,302 participants;
(4) Ouchi and Le Rouzic, 2002, covering 126 World Bank staff learning activities assessed by 2,106 participants;
(5) Quizon, Chard and Lockheed, 2004, covering 192 WBI client learning activities assessed by 794 participants, with the analysis focusing on a subset of 52 activities having detailed activity features data that were assessed by 405 participants.

Because these studies use many different indicators and measures, Table 3 does not attempt to summarize the results quantitatively. Each of the cited meta-analyses provides extensive documentation of both analytic methods and results.

Four dimensions of course quality emerged as consistently associated with positive immediate or mid-term course outcomes: the duration of the course, a course that was professionally designed, a course that required the active involvement of participants, and follow-up. None of these are particularly surprising. Time for learning is one of the most important elements of all types of training, formal and informal. Professional design is associated with courses that were well thought out in advance, rather than hastily put together. And “action learning” or “active learning” with participants encouraged to develop “action plans” to implement what was learned in the course was a positive feature for both staff and client learning. Follow-up ensures that participants are supported while they attempt to implement what they learned.

Dimensions of the course related to “tailoring” it to fit the participants were also positively associated with outcomes: alignment of the course to the participant’s job, homogeneity of participants, and alignment with a country focus. Again, these results are not surprising, given that the closer match between the content of a course and the needs of adult learners, the higher the probability that the learners will benefit.

Conclusion

The introduction of more science-based methods into the evaluation of staff and client learning and capacity enhancement programs at the World Bank has contributed to program improvements in two ways. First, it has allowed for the direct assessment of program effectiveness, whereby attribution of change to the program is plausible. While not all programs were found to boost all indicators of performance, indicators more proximately associated with the program (learning and team products) were sensitive to the intervention effects. Second, the use of multivariate analytic techniques has provided evidence of the comparative effectiveness of various program features, whereby program improvement based

on evidence is possible. Early indications of effective program features were shared with program designers, and these have begun to yield improvements in the design of these learning programs[12]. It remains to be seen whether such improvements result in greater development impact.

References[13]

Bardini, M., Gunnarsson, V., Manjieva, E. M. & Narozhnaya, Y. (2003). *The impact of WBI activities, FY01-02, on participants from Russia: A baseline assessment (WBI Evaluation Studies EG-04-81)*. Washington, D. C.: World Bank Institute.

Bardini, M., Gunnarsson, V. & Palmisano, M. B. (2003). *ACS Network impact evaluation: the "Building your Skills in a Team Based Environment" course. WBI Evaluation Studies EG-04-75*. Washington, D. C.: World Bank Institute.

Bussmann, W., West Meier, M. & Hadorn, A. (2002). *Fiscal decentralization in an era of globalization: an evaluation of the World Bank Institute's Decentralization Program (WBI Evaluation Studies ES-02-54)*. Washington, D. C.: World Bank Institute.

Campbell, D. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand, McNally & Co.

Chard, C. L. & Arango, D. J. (2003). *Annual review of World Bank staff learning, FY03 (WBI Evaluation Studies EG-04-74)*. Washington, D. C.: World Bank Institute.

Cronbach. L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

Denzin, N. K. & Lincoln, Y. S. (Eds.) (1994). *Handbook of qualitative research*. Thousand Oaks, CA: Sage.

Eckert, W.A. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank Institute. *American Journal of Evaluation*, 21 (2), 185-193.

Eckert, W. A., Palmisano, M. B., Gunnarsson, V. & McIntosh-Alberts, M. (2004). *Impact evaluation of sector learning forums (WBI Evaluation Studies EG-04-83)*. Washington, D. C.: World Bank Institute.

Eckert, W. A., Sousa, G. & Gunnarsson, V. (2004). *The impact of WBI activities, FY01-02, on participants from Brazil: A baseline assessment (WBI Evaluation Studies EG-04-82)*. Washington, D. C.: World Bank Institute.

Eckert, W. A., Letiecq, B. L. & Palmisano, M. B. (2003). *PREM Network* Washington, D. C.: World Bank Institute.

Khattri, N., Quizon, J. B., Bardini, M., Eckert, W. A., Prom-Jackson, S., Zia, H, West Meiers, M., Shi, Z., Palmisano, M. B., Novolijov, R., Jha, S., & Nikitin, D. (2002). *Impact evaluation*

[12] For example, the share of WBI activities with "action planning" has increased from 38 percent in FY00-FY01 to 56 percent in FY04.

[13] All WBI reports cited in this paper can be accessed from: http://web.worldbank.org/WBSITE/EXTERNAL/WBI/0,,contentMDK:20252874~menuPK:591798~pagePK:209023~piPK:335094~theSitePK:213799,00.html

of WBI client programs, FY00-01 (WBI Evaluation Studies EG-03-63). Washington, D. C.: World Bank Institute.

Khattri, N., Bachrach, P. & Jiang. T. (2003). *The impact of WBI activities, FY01-02, on participants from Sri Lanka: A baseline assessment (WBI Evaluation Studies EG-04-73)*. Washington, D. C.: World Bank Institute.

Khattri, N., Shi, Z., Palmisano, M. B. & Echternacht, G. (2003). *"Trust Fund Learning and Accreditation" program (TLAP): A review of exam quality and early evaluation results (WBI Evaluation Studies EG-04-71)*. Washington, D. C.: World Bank Institute.

Khattri, N., Shi, Z., Palmisano, M. B. & Fein, M. (2002). *"Trust Fund Learning and Accreditation Program" pilot formative evaluation (WBI Evaluation Studies EG-03-65)*. Washington, D. C.: World Bank Institute.

Kremer, M., Angrist, J., Bettinger, E., Bloom, E. & King, E. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review*, 92(5), 1535-1558. (NBER Working Paper 8343, 2001)

Le Rouzic, V. & Shi, Z. (2003). *Annual review of learning outcomes: WBI client courses, FY02 (WBI Evaluation Studies EG-04-67)*. Washington, D. C.: World Bank Institute.

Le Rouzic, V. (2003). Learning from the *"Introduction to Bank Operations" course, FY03 (WBI Evaluation Studies EG-04-69)*. Washington, D. C.: World Bank Institute.

Liu, C. (2003). *Impact evaluation of "Introduction to Bank Operations" (WBI Evaluation Studies EG-04-76)*. Washington, D. C.: World Bank Institute.

Liu, C., Jha, S. & Stephen Van Praet (2004). *Impact evaluation of WBI Sectoral and Tehmatic Programs, FY02-FY03: Poverty and Growth (WBI Evaluation Studies EG-05-105*). Washington, D. C.: World Bank Institute.

OECD (2002). *DAC Working Party on Aid Evaluation "Glossary of Key Terms in Evaluation and Results Based Management."* Paris: OECD.

Ouchi , F. & Le Rouzic, V. (2002). *Annual review of the quality of formal World Bank staff learning FY02 (WBI Evaluation Studies EG-03-56)*. Washington, D. C.: World Bank Institute.

Ouchi, F., Shi, Z. & Zhou, C. (2001). *The performance of WBI core course training, FY 1999-2001: An evaluation (WBI Evaluation Studies ES-02-55)*. Washington, D. C.: World Bank Institute.

Prom Jackson, S., Cooper, M., Sankofa, B., Martin, B. S., Kategile, W. J., Palmisano, M. B. & Arango, D. (2002). *The quality and impact of the "Individual Coaching"program, 2000-2002 (WBI Evaluation Studies EG-03-57)*. Washington, D. C.: World Bank Institute.

Prom-Jackson, S., Cooper, M., Palmisano, M. B., Latib, M., Rickwood, B. & Arango, D. (2002). *The quality and impact of the "Challenge of Leadership" seminar, 1999-2002 (WBI Evaluation Studies EG-03-59)*. Washington, D. C.: World Bank Institute.

Quizon, J. B., Ouchi, F. & Gunnarsson, V. (2004). *Phase 3 evaluation of the "Multi Sectoral Team Learning (MTL)" program. (WBI Evaluation Studies EG-04-90)*. Washington, D. C.: World Bank Institute.

Quizon, J. B., & Chard, C. L. (2003). *The impact of WBI activities, FY01-02, on participants from Thailand: A baseline assessment (WBI Evaluation Studies EG-04-77)*. Washington, D. C.:

World Bank Institute.

Quizon, J.B., Chard, C. L. & Lockheed, L. (2004). *The effectiveness and usefulness of WBI learning events: An FY01-02 baseline assessment in five WBI focus countries. (WBI Evaluation Studies EG-04-86)*. Washington, D. C.: World Bank Institute.

Rossi, P. H. & Freeman, H. E. (1993). *Evaluation: A systematic approach (5th edition)*. Newbury Park, CA: Sage.

Rubin, D. B. & Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal covariates. *Biometrika*, 79, 797-809.

Scriven, M. (1991). *Evaluation thesaurus (4th ed.)*. Newbury Park, CA: Sage.

Universalia (2001). *Sprite impact study*. Montreal, Quebec.

Zia, H., Al-Sayyid, M. K., Tawila, S. & Gunnarsson, V. (2003). *The impact of WBI activities, FY01-02, on participants from Egypt: A baseline assessment. (WBI Evaluation Studies EG-04-78)*. Washington, D. C.: World Bank Institute.

Zia, H. (2004). *The impact of the "New Manager's Leadership" course, 2003-2004. (WBI Evaluation Studies EG-05-103)*.Washington, D. C.: World Bank Institute.

Annex A: Additional Information on Cases with Counterfactuals

Case	Course Name	Duration	Objective	Total Participants (particip./control)	Outcome indicator(s)
Post-Learning Case I	Building Your Skills in a Team-Based Environment	5 days	Improving interpersonal effectiveness, information seeking, teamwork and cooperation, acting assertively, client orientation, and quality and timeliness of work.	269 (84/87)	Managers' ratings of participants' behavioral skills
Post-learning Case II	Sector Forum (N = 10 forums)	2-7 days	Sharing experience, agreeing on priorities, identifying key issues, improving job performance	3212 (1666/4073)	Managers' ratings of participants' job performance
Post—learning case II	Multi-Sectoral Team Learning	variable	Improving teaming and learning practices, processes and tools; improving the quality of products	50 teams (12/8)	Independent quality ratings of team products
Pre-learning Case I	Introduction to Bank Operations (N = 4 cohorts)	5 days	Increasing knowledge, improving application of knowledge on the job	120 (120/ 300)	Test of knowledge gain, retention of knowledge, use of knowledge and perceived effect of course on perceptions and motivation
Pre-learning Case II	New managers leadership program (N = 2 cohorts)	15 days in 3 5-day modules	Improving management skills, achieving business results, leading change	70 (35/35)	Direct reports' ratings of manager's performance